

ABSTRACT

Provided is a method and an apparatus for controlling quality of service of a virtual machine, relating to the fields of cloud computing, virtualization, and big data. The method includes: storing (S101) an operation request in a target queue corresponding to a layer of a virtual machine cluster when an available token quantity in a token bucket of the layer is unable to meet a target token quantity required for the operation request; wherein tokens of the layer are periodically generated according to an upper limit of quality of service of the layer; deducting (S102) the target token quantity from an updated available token quantity when the available token quantity in the token bucket is updated to meet the target token quantity required for the operation request; and sending (S103) the operation request from the target queue to the layer for processing when deducting the target token quantity successfully.

[Figure No. 1]